14.310x: Data Analysis for Social Scientists
Special Distributions, the Sample Mean, and the Central Limit Theorem

Welcome to your fifth homework assignment! You will have about one week to work through the assignment. We encourage you to get an early start, particularly if you still feel you need more experience using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced.

Some of the questions we are asking are not easily solvable using math so we recommend you to use your R knowledge and the content of previous homework assignments to find numeric solutions.

Question 1
A manufacturer receives a shipment of 100 parts from a vendor. The shipment will be unacceptable if more than five of the parts are defective. The manufacturer is going to randomly select K parts from the shipment for inspection, and the shipment will be accepted if no defective parts are found in the sample.

How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable shipment is less than 0.1?
Hint: We recommend using R to plug in different values of K.

- 42
- 22
- 32
- 12

Question 2
Now suppose that the manufacturer decides to accept the shipment if there is at most one defective part in the sample. How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable shipment is less than 0.1? As above, a shipment is unacceptable if there are more than 5 defective parts.

Question 3
A man with n keys wants to open his door and tries the keys at random. Exactly one key will open the door.
What is the expected value of the number of trials needed to open the door if unsuccessful keys are not eliminated for further selections?

- n
- $n-2$
- $\frac{n-1}{n+1}$
- $n-1$

Question 4
Let the number of chocolate chips in a certain type of cookie have a Poisson distribution. We want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than 0.99. For which of the following values of the mean of the distribution is this condition assured? (Please select all that apply!)

Hint: You may wish to try different values in R when solving this problem if you have trouble solving the relevant equations.

- ☐ 6
- ☐ 7
- ☐ 8
- ☐ 9

You decide to move out of your college's dorms and get an apartment, and you want to discuss the budget with your roommate. You know that your monthly grocery bill G will depend on a number of factors, such as whether you are too busy to cook, whether you invite guests for meals frequently, how many special holiday meals you will cook, etc. In particular, G will have an approximate normal distribution with a variance of 2500 and a mean:

$$\mu = 300 + 10M - 100B + 50H$$

where M is the number of meals to which you invite guests, and $\mathbb{E}[M] = 8$. B is a measure for how busy you are with 14.310x problem sets and assume it is $U[0,1]$. H is a variable that takes on the value 1 for holiday months of November, December, and January and 0 otherwise.

Question 5
What is the mean of G in November, where $M = 10$ and $B = 0.5$?

Question 6
For a month chosen at random what is $\mathbb{E}[G]$? (Select all that apply)

- ☐ $\mathbb{E}[300 + 10M - 100B + 50]$
- ☐ $\mathbb{E}[300 + 10M - 100B]$
- ☐ $300 + 10\mathbb{E}[M] - 100\mathbb{E}[B] + 50 * \mathbb{E}[H]$
- ☐ $312.5 + 10\mathbb{E}[M] - 100\mathbb{E}[B]$
- ☐ $\mathbb{E}\left[300 + 10M - 100B + 50 * \frac{3}{4}\right]$
- ☐ $\mathbb{E}\left[300 + 10M - 100B + 50 * \frac{1}{4}\right]$

Question 7
What is $E(G)$?

Now we are going to perform some simulations in R. We are going to follow Sara's example in the lecture where we imagine a case where the x_i follow a uniform distribution between 0 and θ ($U[0, \theta]$), and two researchers are trying to figure out the value of θ. (We will set $\theta = 5$). We are going to simulate different random samples from this distribution with a sample size of 100

observations each. These samples will be available to the two researchers, and we are going to plot how $\hat{\theta}$ is distributed for different estimators.

There are two types of researchers in this world. Researcher A uses as an estimator for $\theta, \hat{\theta}_A = 2 * \bar{x}$ where $\bar{x}$ corresponds to the sample mean of the sample he receives from us. Researcher B uses as an estimator $\hat{\theta}_B = 2 * median(x)$ where $median(x)$ corresponds to the median of the sample he receives from us.

We have provided you this R code that has some information missing in case you need help for this exercise.

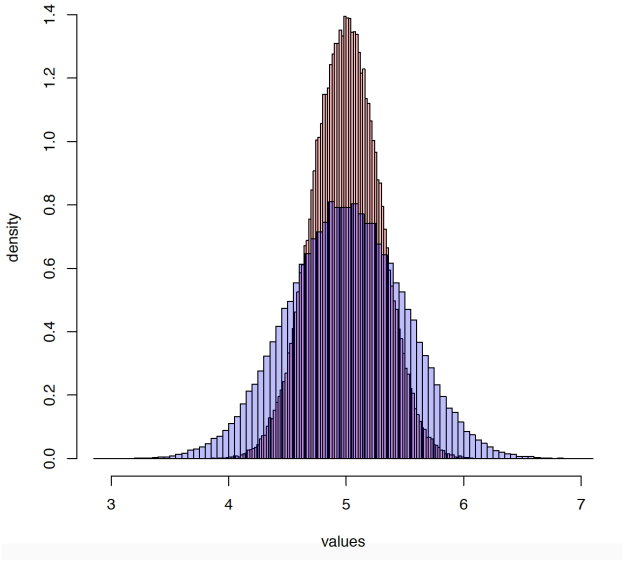
Question 8
What would be the mean of this distribution of $\bar{x}$?

Question 9
What would be the variance of the distribution of $\hat{\theta}_A$? Please enter the numerical value of the variance.

Note: Please review our guidelines on precision regarding rounding answers here.

If you haven't already, please fill in the R code provided earlier.

We have run our simulations, simulating 100,000 different samples of size 100. We have provided 200,000 researchers (A and B), each with one of these samples. They have sent us their estimators for $\hat{\theta}$. The following plot shows a histogram of their estimators (Figure 1).



Question 10
Does the blue histogram correspond to the estimator of researcher A or researcher B?

- Researcher A
- Researcher B

Question 11

Since both of the estimators are centered around the real value of the parameter θ, you should use the estimator with the lowest variance. Which estimator should you use?

- $\hat{\theta}_A$
- $\hat{\theta}_B$

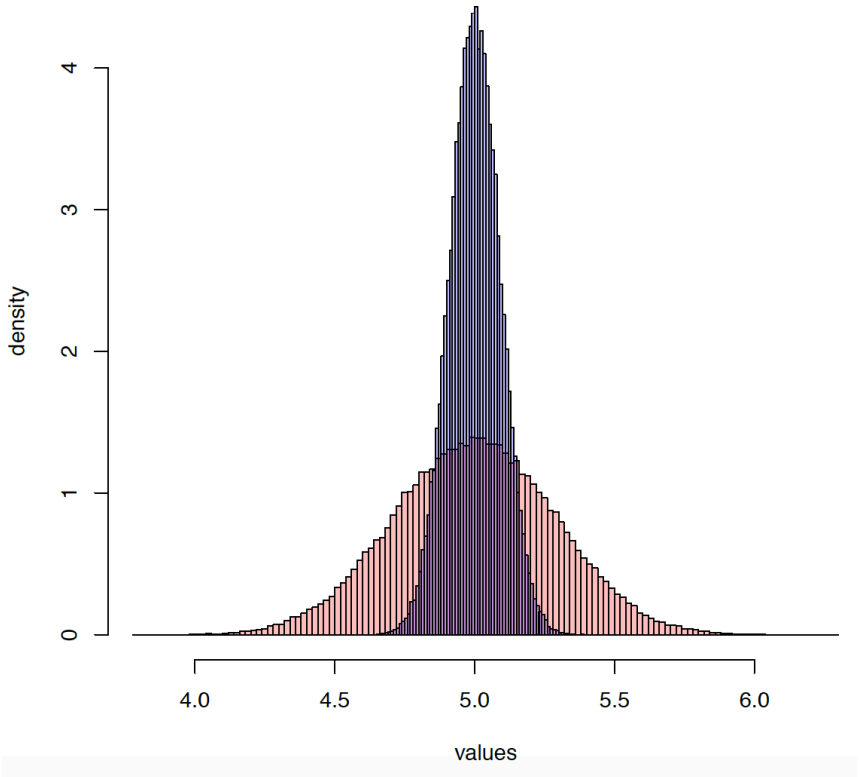
Question 12

Now, let's increase the sample size to 1000. As an exercise try to use the provided code to code this yourself in R. What would be the new variance of the estimator $\hat{\theta}_A$?

Note: Please review our guidelines on precision regarding rounding answers here.

Question 13

The following figure shows the distribution for $\hat{\theta}_A$ for $n = 100$ and $n = 1000$



Does the blue histogram correspond to a sample size of 100 or of 1000?

- 100
- 1000